Hybrid Metric-Topological-Semantic Mapping in Dynamic Environments

Romain Drouilly, Patrick Rives, Benoit Morisset

Hybrid Metric-Topological-Semantic Mapping in Dynamic Environments

Romain Drouilly [1,2], Patrick Rives[1], Benoit Morisset[2]

***Abstract*— Mapping evolving environments requires an update mechanism to efficiently deal with dynamic objects. In this context, we propose a new approach to update maps pertaining to large-scale dynamic environments with semantics. While previous works mainly rely on large amount of observations, the proposed framework is able to build a stable representation with only two observations of the environment. To do this, scene understanding is used to detect dynamic objects and to recover the labels of the occluded parts of the scene through an inference process which takes into account both spatial context and a class occlusion model. Our method was evaluated on a database acquired at two different times with an interval of three years in a large dynamic outdoor environment. The results point out the ability to retrieve the hidden classes with a precision score of 0.98. The performances in term of localisation are also improved.**

I. INTRODUCTION

Lifelong mapping has received an increasing amount of attention during last years, largely motivated by the growing need to integrate robots into the real world wherein dynamic objects constantly change the appearance of the scene. A mobile robot evolving in such a dynamic world should not only be able to build a map of the observed environment at a specific moment, but also to maintain this map consistent over a long period of time. It has to deal with dynamic changes that can cause the navigation process to fail. However updating the map is particularly challenging in large-scale environments. To identify changes, robots have to keep a memory of the previous states of the environment and the more dynamic it is, the higher will be the number of states to manage and the more computationally intensive will be the updating process. Mapping large-scale dynamic environments is then particularly difficult as the map size can be arbitrary large. Additionally, mapping many times the whole environment is not always possible or convenient and we could take advantages of methods using only a small number of observations. The idea exploited in this paper is to use scene understanding to retrieve a stable world model with only two acquisition sequences.

Previous mapping strategies developed for dynamic environments can be grouped in a few categories. The first group concerns methods that remove dynamic objects in order to achieve a stable representation [1]. A strategy to identify dynamic objects in the scene and map them in a separate occupancy grid is proposed in [2]. Those methods requires to identify moving objects which make them more suited for fast dynamic changes. Another approach that does not require to explicitly identify moving objects is presented in [3]. It consist in maintaining several maps acquired at different time scale and to select the most adapted at a given time for navigation by checking its consistency with the current observations. The main drawback of this approach is the large amount of data needed, five times as big as a single map, which prevents its use for large-scale environments where the size of the map is already a problem. A third group of methods assumes that mapping is a never-ending process and continuously update data. The biologically inspired algorithm RATSLAM is used in [4] to perform persistent mapping at the cost of an increasing map size. Similarly, an hybrid metric-topological map is proposed in [5] where a model based on the human memory is used to update a feature-based description of spherical views constituting the map. The update process consists in adding information about stable features and in removing features that no longer exist in spherical images. Once again, the major issue is to deal with the large amount of data. A last approach consists in transposing the problem in another space where changes

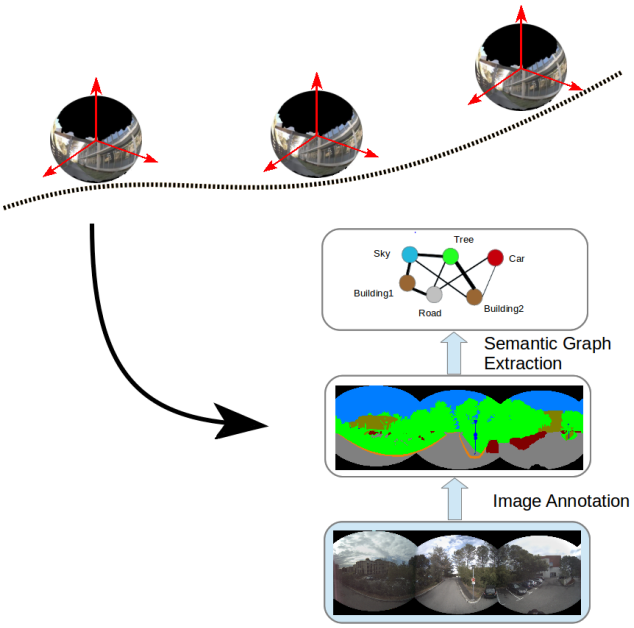


Fig. 1. Map Structure. Spherical images augmented with depth are captured while the robot explore its environment. Then each spherical image is automatically annotated. Finally a *semantic graph* is extracted from the annotated image

*This work was supported by ECA Robotics

[1]Authors are with tem Lagadic at INRIA Mditerrane, France romain.drouilly@inria.fr, patrick.rives@inria.fr

[2]Authors are with ECA Robotics, bmo@eca.fr

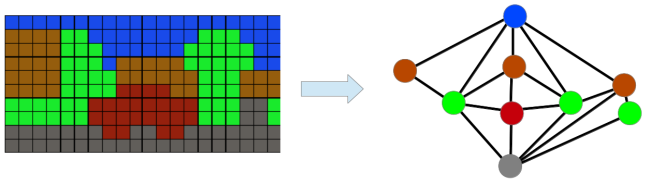


Fig. 2. Example of semantic graph extracted from an annotated image

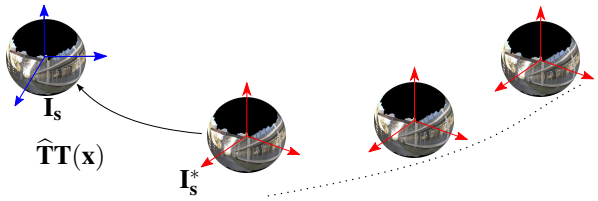


Fig. 3. Illustration of the registration process. The current sphere is in blue and the reference spheres are in red.

are easier to handle and to model. An example, introduced in [6], is the use of spectral analysis to model dynamics of the environment. However identification of dynamic behavior requires once again, large amount of observations which is the the main problem of most of the previous approaches, making them unsuitable for large-scale environments where the amount of data for a single mapping session is already significant.

In this paper we propose a framework adapted to large scale outdoor environments that relies on the hybrid Metric-Topological-Semantic (MTS) map, introduced in our previous work[7]. We show how our mapping approach can naturally handle changes in the environment for both mapping and localisation through the use of semantic maps. Instead of updating directly the low-level layers of the map, we update semantic data and generalise changes observed over space to changes over time using the ergodic assumption.

II. PROPOSED APPROACH

In this work we propose a new scheme to update the map of large-scale dynamic environments using a stable representation based on semantic information. While many mapping systems rely on a representation based on low-level features descriptors, our method only requires semantic description of a set of reference images [7], [8]. Consequently, the changes occurring in the scene will be coded in the semantic layer of the map directly. In the presented approach, instead of stacking all perceptions at a very high computational cost, a compact representation of changes is built. More precisely, the set of possible classes is split in two groups, namely dynamic and static classes. Then, changes are modeled in terms of static classes occlusions due to dynamic objects. This model takes the form of a probability distribution that encodes the risk that a given dynamic class occludes a static class. It is build from the observation of occlusions over space and time and is used in conjunction with contextual information to infer semantics of occluded areas. For the sake of completeness, we briefly review in the next two sections our MTS map structure, illustrated at figure 1, and our localisation strategy, previously introduced in [7]. Then the proposed map update framework is detailed.

A. MTS Map Structure

Our map consists of a set of spherical RGB images augmented by depth and semantic data, as illustrated at fig 1. Those images, so-called reference images, are multi-layers local submaps of the environment perceived from a particular viewpoint. The metric layer is built from data acquired with a multi-camera stereovision system previously described in [9]. The sensor consists in two superimposed rings of three cameras allowing to capture both photometric and geometric data over a 360^o field of view. At each pixel is attributed a semantic label thanks to a two steps process as presented in [7]. Firstly, *Random Forest* are used to classify SIFT-based descriptors extracted densely from images. Then a *Fully Connected Conditional Random Field* is used to model neighborhood and an efficient inference method [10] allows us to correct the labels over spatial context. From these annotated images are extracted *semantic graphs*, illustrated at figure 2. Let A_i be a group of contiguous pixels with the same label in the annotated image, called a *semantic area*. A *semantic graphs* is denoted as $g = \{A, E\}$ where A is the set of *semantic areas* and E the set of edges encoding their adjacency in the annotated image. Each $A_i \in A$ is characterized by a fitted ellipsoid envelop $f_i = \{x_i, y_i, h_i, w_i, \alpha_i\}$ where (x, y) is the position of the ellipse, (h_i, w_i, α_i) its main axis and orientation respectively.

Semantic graphs are local powerful representations of the environment as they encode both scene structure and high-level description of the context in a very compact way. The localisation strategy strongly relies on those graphs. At a larger scale, all submaps are positioned in the scene thanks to a dense visual odometry method presented in [11] and constitute a global graph of the environment.

B. Localization in MTS Map

Localization in MTS map is a coarse-to-fine two steps process. For a given image of the current scene, a similar submap is efficiently retrieved in the global graph using semantics. The semantic graph g_{cur} extracted from the current annotated image is compared to the N semantic graphs g_i of the map using an *Interpretation Tree*. This algorithm allows to compare efficiently graphs using both nodes appearance and neighbors. Each nodes of the semantic graphs with similar labels are matched two by two using unary constraints, capturing their intrinsic properties (h_i, w_i, α_i), and pairwise constraints, capturing the context consisting of the nearest neighbors in the graph.

Matching graphs allows to compute a *similarity score*, denoted as σ, between two semantic graphs G_1 and G_2. It is measured as follows:

$$\sigma(G_1, G_2) = \exp^{1-\frac{N}{N_m}} \tag{1}$$

where N_m is the number of nodes matched between the two graphs denoted as $A_{12} = A_{G1} \bigcap A_{G_2}$ and N the total number

of nodes in the current semantic graph. The submap with the highest score σ corresponds to the most probable closest location.

Once the submap corresponding to the closest position is retrieved, a dense registration method between the submap and the current spherical image, described in [12], is applied to refine the pose estimate locally (see figure 3). Pose estimation between a current spherical image I_s and a retrieved spherical image I_s^* is done using robust minimization techniques. Following the formulation of [13], the cost function for optimising intensity errors between spheres $\{I_s, I_s^*\}$ is given as:

$$\mathfrak{F}_I = \frac{1}{2}\sum_i^k \Psi_{hub} \left\| I_s\big(\omega(\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x}); P_i)\big) - I_s^*\big(\omega(\mathbf{I}; P_i^*)\big) \right\|^2, \quad (2)$$

where $\omega(.)$ is the warping function that projects a 3-D point P_i given a pose $\mathbf{T}$ onto the sphere. The pose $\widehat{\mathbf{T}}\mathbf{T}(\mathbf{x})$ is an approximation of the true transformation $\mathbf{T}(\tilde{\mathbf{x}})$ and Ψ_{hub} is a robust weighting function on the error given by Huber's M-estimator [14].

C. Updating the Maps

They are two main changes occurring in a dynamic environment. Those due to dynamic objects and those caused by illumination changes (day/night). As we are focused on life-long mapping, we consider here only the changes in the scene due to occlusions caused by dynamic objects. This choice is motivated by the fact that, in a semantic approach, robustness to illumination changes should be more easily handled by the classification process, which is beyond the scope of this article.

A dynamic class, denoted as C_D occludes a static class C_S by changing the label associated to the corresponding pixels in the image. The objective is to identify the parts of the images corresponding to dynamic classes and to retrieve the static class occluded by the dynamic object in order to achieve a stable representation. Let us consider two different cases depending if the scene is re-observed or not.

1) Map updates in re-observed areas: When the robot navigates its environment, it can observe several times the same place. Dynamic objects may have moved and then previously occluded areas can be observed. These observations are used to build more stable semantic representation by replacing previously occluded areas in the reference annotated image, by the labels provided by the new observation. To do this, the pose between the two images is computed using dense matching techniques. Then a semantic warping function is used to project labels of the new observation onto the reference annotated image.

Let I_s^* be the reference annotated image of size $m \times n$ from which we want to compute the stable representation. A pixel in I_s^* is identified by its position $p^* = (u, v)$, where $u \in [0, m[$ and $v \in [0, n[$. A 3D point in the euclidean space is defined as $P^* = \{p^*, Z, l\}$ where $Z \in \mathbb{R}^+$ is the depth expressed in the image reference frame and $l \in \mathscr{L}_S$ the associated label. Let I_i be another observation of the same scene view from the position $\mathbf{T}(x)_i$, where $T(x)_i$ is expressed in the reference frame of I_s^*.

It is possible to synthesize a new annotated image, denoted as I_i' from the labels $L(p)$ of I_i at the position of the reference image using the warping function:

$$p^{\omega} = \omega(\mathbf{T}(x)_i; Z, l, p^*) \quad (3)$$

where $\omega(.)$ function lifts the pixel p^* from the reference image to the new observed image using the rigid transform $T(x)_i$ followed by a spheric projection. The projected point does not correspond exactly to the pixels position so a closest neighbor interpolation is used to select the corresponding label. Finally, if a given pixel p_i is associated to a class $C_j \in C_D$ in I_s^* a class $C_k \in C_S$ in I_i', its class is set to C_k.

2) Map update in unobserved areas: The warping function allows a partial update of the map where additional observation informs about the underlying static classes. But some areas may remain unobserved if dynamic objects occlude the same part of the scene for the two mapping sequences.

To deal with these areas, we need to do *semantic inpainting*: unobserved areas are treated as holes that have to be filled with static class labels. A model of occlusions occurring in the scene is computed to infer the label of the pixels that remain occluded by dynamic objects. It relies on the *ergodic assumption* which states that the average behavior of dynamic objects over the time is essentially the same as the average behavior of dynamic objects over the space. More precisely, this assumption state that we can generalize occlusions observed over the mapping sequences to unobserved areas which remain occluded in other parts of the map. Practically, we compute a model that describes which static class is likely to be occluded by a given dynamic class. For example the dynamic class "pedestrian" is more likely to occlude the static class "side-walk" than "sky" as the class "car" is likely to occlude "road". This model takes the form of a probability distribution of the existence of an underlying static class C_i given the observation of a dynamic class C_j:

$$P(C_i|C_j) = \frac{O(C_i, C_j)}{N} \quad (4)$$

where N is the total number of pixels initially associated to a dynamic label and $O(C_i, C_j)$ the number of pixels initially labelled as C_j and corrected to C_i using the warping function. This model is computed thanks to observations made over several acquisition sequences. It is not necessary to remember specific observations but only the model which is extremely compact.

However this model is not sufficient to correctly estimate the occluded classes because it only takes into account statistics over time. It is necessary to take into account the spatial context to estimate the probability that a given static class is occluded by a dynamic class. For example if a dynamic object of type "car" is mainly surrounded by the class "building", it is more likely that the car occludes a building than a road even if roads are usually the most probable occluded class.

To take into account the context, the semantic graph associated with the annotated image is used. For a dynamic area in the image, the semantic graph gives the adjacent semantic areas, constituting the neighbors, denoted as $\mathcal{N}$. Each node n_i in a semantic graph is characterized by a fitted ellipsoid f_i to describe its shape [7] which parameters are presented at section II-A.

To model the probability of associating a static label to a pixel $p = (x_p, y_p)$, a Gaussian function is associated to each neighbor node $n_i \in \mathcal{N}$. It takes the general form:

$$F_i(x_p, y_p) = A_i exp^{(-a(x_p - x_i)^2 + 2b(x_p - x_i)(y_p - y_i) + c(y_p - y_i)^2)} \quad (5)$$

where A_i is the amplitude set as $P(C_i|C_j)$, (x_i, y_i) the position of the area in the image, and where:

$$a = \frac{cos^2\theta}{2\sigma_x^2} + \frac{sin^2\theta}{2\sigma_y^2} \quad (6)$$

$$b = \frac{sin(2\theta)}{4\sigma_x^2} - \frac{sin(2\theta)}{4\sigma_y^2} \quad (7)$$

$$c = \frac{sin^2\theta}{2\sigma_x^2} + \frac{cos^2\theta}{2\sigma_y^2} \quad (8)$$

with $\sigma_x = h_i$,$\sigma_y = w_i$ and $\theta = \alpha_i$.

Then for each pixel of the area requiring a new label, the most probable label is computed as follows:

$$L(C_i) = \max_{i \in \mathcal{N}}(F_i(x_p, y_p)) \quad (9)$$

where $L(.)$ stands for the likelihood.

Using the proposed approach, it is possible to update the map by exploiting both the spatial context and the knowledge acquired along robot's experience, resulting in a robust and stable representation of the environment. Conversely to many other approaches, we do not require to consider a large set of observations but only a simple and compact model of occlusions.

III. EXPERIMENTS

Our framework has been tested in two ways. First, the correctness of the class prediction in occluded parts of the scene has been evaluated by making predictions in areas where observations of static labels are accessible and used as ground truth. Then, the usefulness of the approach for localization is evaluated by comparing similarity scores of images taken at the same place but at different moment with and without updating data. All experiments were performed using an Intel i7-3840QM CPU at 2.80GHz. All programs are single-threaded.

The two experiments are realized with a challenging dataset modeling an outdoor environment with forest and building areas at the INRIA Sophia-Antipolis campus. It is composed of hundreds of high resolution[1] spherical images taken along a 1.6km pathway. Two sequences have been acquired with our multi-camera stereovision system on the same pathway at two different time-scales with an interval of three years. The automatic annotation of images produces 9 classes: tree, sky, road, sign, sidewalk, ground sign, building, car, other. The scene parsing stage produces a not perfect labelling, achieving 82% of correctly labelled pixels (see more results in [7]). It is important to notice that learning is done only on images extracted from the first sequence. Then changes in illumination are managed by the classification stage only.

[1]The full resolution is 2048x665 but we use 1024x333 resolution for classification

TABLE I
OCCLUSION MODEL

Class	Associated Probability
Sky	0
Building	0.04
Road	0.76
Sidewalk	0.08
Tree	0.11
Signs	0
Ground Signs	0

TABLE II
INFERENCE RESULTS

Class	Score
Building	0.96
Road	0.98
Sidewalk	0.99
Tree	0.97
Global	0.98

A. Predictions Correctness

To estimate the correctness of the predictions, we report two measures. The first one is the global precision of the inference process. The global precision is given by the number of pixels correctly associated to the class C_i over the total number of pixels labelled as C_i, denoted as S_{global}. But this measure alone is not significant as some classes representing small objects can be ignored without significant decrease of performance. Then we also report by class unweighed precision score, denoted as S_{class}. The occlusion model computed from experiments is reported at table I. The results of inference process are reported in table II and illustrated at figure 4.

As expected the occlusion model encodes the common fact that cars are more likely to appear on road, in front of trees or buildings than in the sky. Results presented in table II for classes with non null probabilities, show that the inference process is very efficient. Almost all pixels are associated with the correct class. The remaining pixels correspond to slight changes in border position. These very good results demonstrate the efficiency of our approach to infer semantics in occluded parts of the scene. These good results can be explained by two facts. First, taking into account both spatial context and temporal changes allows to build a very robust model of the world. The ergodic assumption is a very efficient way to compensate for the small number of observations, only two here. The second point is that static

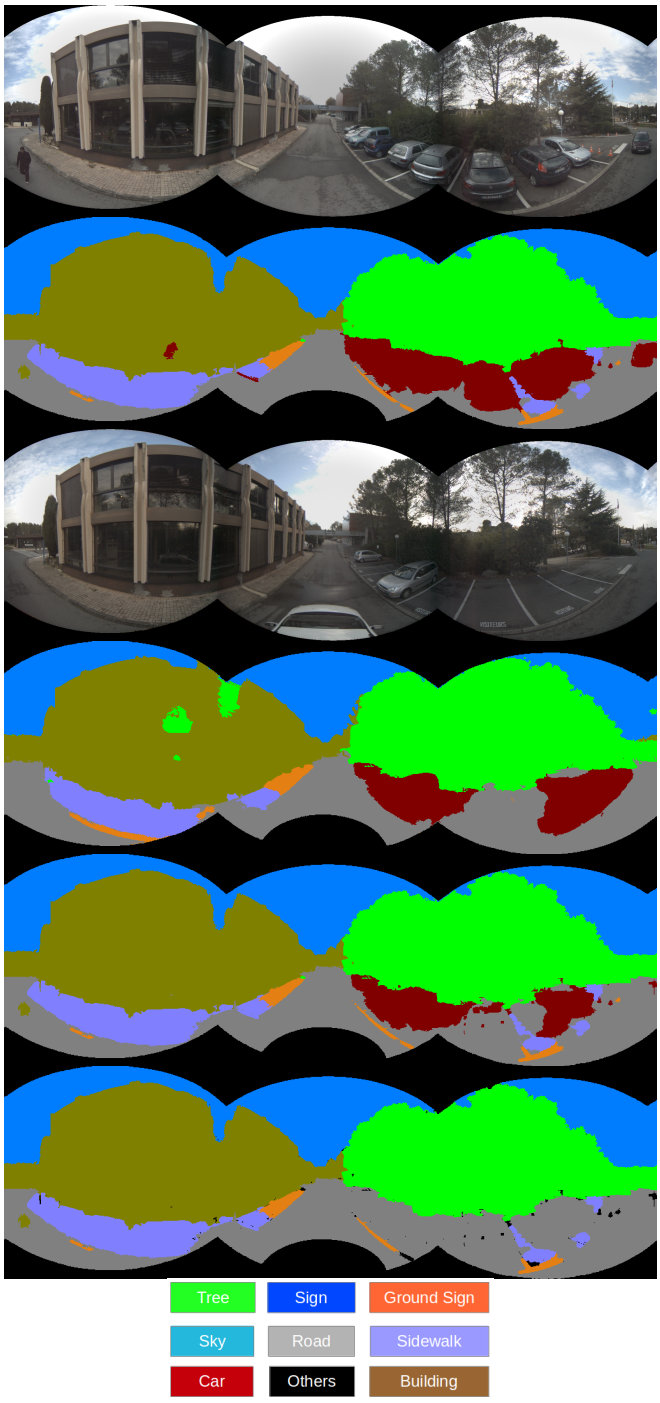


Fig. 4. Example of Inference Results. *From top to bottom*: Initial image and scene parsing results; Image at the same place, three years later and scene parsing results; Initial image updated by new information; Initial image updated by inference

classes are most of the time only partially occluded. Then they can be taken into account to fill the remaining dynamic areas. If a class were completely occluded, it would have been impossible to predict its presence with our model. However, there is a low probability that this happen because the remaining occluded areas after several observations are rather small. Then the probability to hide completely an object is small too.

Another interesting point is illustrated in the figure 4. The four top images show two pictures of the same place taken at a three years interval. The cars are not at the same place. However the classification stage makes an error predicting cars at a place where no car is observed in the second image. The updating process allows to correct this error, increasing the precision of scene parsing by the same way.

Finally it show that even with a small number of acquisition sequences, we are able to achieve a stable representation of the scene. Using both contextual and temporal information to discover underlying static classes.

B. Localization

The second way to estimate the usefulness of our approach is to measure how it affects localization. As mentioned earlier, we use a semantic based localization process to select the local submap corresponding to the current view. Re-localization could be challenging in a dynamic scene where objects changes scene appearance over time. This experiment shows how our approach could be useful. To do this we compare the similarity score, given by equation 1, for images of the same place taken at different timescales with and without updating data with warping and inference. One hundred images taken in the most dynamic areas are used for this test. Results are presented in table 6.

As expected, results show that our approach allows to increase the similarity between images taken at the same position over the time. The matching process relies on semantic areas shapes which are modified by occlusion. Then dealing with large occlusions allows a significant increase in matching performances. The proposed updating process allows to increase the robustness of localization by better characterizing objects shapes. The figure 5 gives an illustration of the results. Images at the left are not updated. Changes in cars positions modify the shape of surrounding semantic areas so that only 7 semantic areas over 14 are matched resulting in the similarity score $\sigma = 0.42$. Using the updated images to the right, 8 objects over 11 are matched giving the similarity score $\sigma = 0.65$. Our approach is then a way to improve the navigation by increasing the robustness of representation with respect to dynamic objects.

IV. Conclusion

We have presented a new approach to large-scale dynamic environments mapping using semantics. Our approach detects changes observed over time at the object level and builds a model of occlusions caused by dynamic objects. Then this model and the spatial context are exploited to recover labels of hidden parts of the scene through an inference process which generalize observations across the map. Conversely to many other works, updating the map does not require large amount of observations what makes it well suited for large-scale environments where the size of a single map is already a problem. The framework has been tested on a real-world dataset acquired at two different timescales with a three years interval and has shown very promising

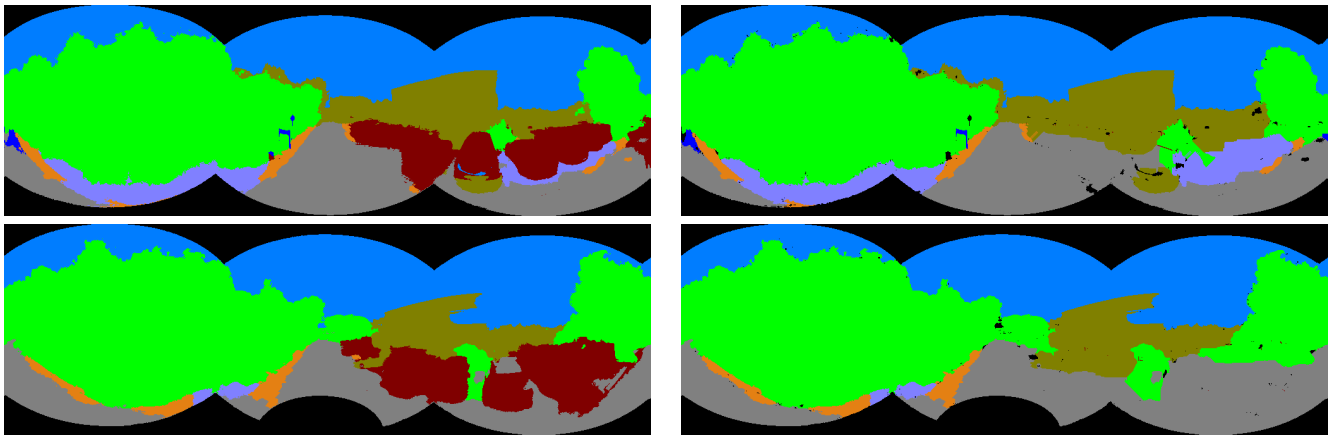


Fig. 5. Examples of images of the two datasets with and without upadted data. Top left: image of the first sequence; Bottom left: corresponding image in the second sequence. Top right: updated image of the first sequence; bottom right: updated image of the second sequence

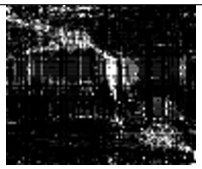
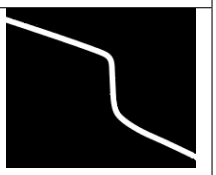
Measure	Graph	Updated Graph	Ground truth
			
σ	0.42	0.65	-

Fig. 6. Similarity matrix with and without updating graphs. It measures the similitude measured between images taken at initial time (row) and three years after (column). The higher the similitude the higher the pixel value. The similarity matrix is better with updated graphs, showing the usefullness of our approach.

results. The precision score of the inference process reach 0.98 and the localisation has been shown to be more robust.

However some improvements can be made to our approach. for example the assumption that classes labelled as dynamic have to be remove is generally true. But it is possible that some instances of these classes correspond to static objects, like an abandoned car. In these cases, removing them from the map introduces errors. Our approach can benefit from integrating these cases in the model.

ACKNOWLEDGMENT

The work presented in this paper was supported by the French Fundation of technological Research under the grant CIFRE N2012/0067.

REFERENCES

[1] D. S. D. Hähnel and W. Burgard, "Mobile robot mapping in a populated environments," n. . p. .-. Advanced Robotics, vol. 17, Ed.

[2] D. F. Wolf and G. S. Sukhatme, "Mobile robot simultaneous localization and mapping in dynamic environments," *Autonomous Robots*, vol. 19, no. 1, pp. 53–65, 2005.

[3] P. Biber and T. Duckett, "Experimental analysis of sample-based maps for long-term slam," 2008.

[4] M. Milford and G. Wyeth, "Persistent navigation and mapping using a biologically inspired slam system," *The International Journal of Robotics Research*, vol. 29, no. 9, pp. 1131–1153, 2010.

[5] F. Dayoub, G. Cielniak, and T. Duckett, "Long-term experiments with an adaptive spherical view representation for navigation in changing environments," *Robotics and Autonomous Systems*, vol. 59, no. 5, pp. 285 – 295, 2011, special Issue {ECMR} 2009.

[6] T. Krajník, J. Fentanes, G. Cielniak, C. Dondrup, and T. Duckett, "Spectral analysis for long-term robotic mapping," in *International Conference on Robotics and Automation (ICRA)*, 2014.

[7] R. Drouilly, P. Rives, and B. Morisset, "Fast hybrid relocation in large scale metric-topologic-semantic map," in *IROS 2014*, 2014.

[8] ——, "Semantic representation for navigation in large-scale environments," in *ICRA 2015*, 2015.

[9] M. Meilland, A. Comport, and P. Rives, "Dense omnidirectional rgb-d mapping of large scale outdoor environments for real-time localisation and autonomous navigation," *Journal of Field Robotics, "Special Issue on Ground Robots Operating in dynamic, unstructured and large-scale outdoor environments"*, vol. 32, no. 4, pp. 474–503, June 2015.

[10] P. Krahenbuhl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems 24*, J. Shawe-taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., 2011, pp. 109–117.

[11] T. Gokhool, R. Martins, P. Rives, and N. Despr, "A compact spherical rgbd keyframe-based representation," in *IEEE Int. Conf. on Robotics and Automation, ICRA'15*, Seattle, WA, May 2015, pp. 4273–4278.

[12] M. Meilland, A. I. Comport, and P. Rives, "A spherical robot-centered representation for urban navigation." in *IROS*. IEEE, 2010.

[13] T. Gokhool, M. Meilland, P. Rives, and E.-F. Moral, "A Dense Map Building Approach from Spherical RGBD Images," in *Int. Conf. on Computer Vision Theory and Applications, VISAPP*, Lisbon, Portugal, January 2014.

[14] P. Huber., "Robust statistics," . New york, Wiley, Ed.